

Context Sensitivity of the Force Concept Inventory

An Honors Thesis submitted in partial fulfillment
of the requirements of Honors Studies in Physics

By
Heather A. Griffin

Spring 2004
Physics
J. William Fulbright College of Arts and Sciences
University of Arkansas

Acknowledgements

This work received support from the Honors College in the form of a grant. Thank you to all the people who made that possible by writing recommendations and providing comments during the application process. My deepest gratitude goes out to Dr. John Stewart for all the guidance and patience exhibited while working with me and for always having the time to help. A very special thanks to Dr. Gay Stewart not only for the help with administrative tasks related to this project, but also for being the most attentive and caring academic advisor over the last four years that a student could ask for.

Table of Contents

Acknowledgements.....	2
Abstract.....	4
Chapter 1: Introduction.....	4
Chapter 2: Data Collection.....	11
Chapter 3: Context Sensitivity.....	13
Chapter 4: Correlation.....	17
Chapter 5: Response Distribution.....	19
Chapter 6: Random Model.....	21
Chapter 7: Uncertain Model.....	23
Chapter 8: Testing Orders.....	26
Chapter 9: Summary and Conclusions.....	29
Bibliography.....	32
Appendix.....	34

Abstract

The results of comparing the Force Concept Inventory and a context-modified test are given for 654 students enrolled in University Physics I at the University of Arkansas. A statistical t-test concludes that the context changes did have a significant effect on student responses, but the overall student performance is not affected by the context changes. Probabilities are calculated to determine whether a majority of the wrong answers are the result of using an incorrect model, random error, or uncertainty. An F-test concludes the effect of testing order is significant, implying that the first test affects responses given on the second test.

Chapter 1: Introduction

Reformation of undergraduate education in physics as well as other fields of science and engineering has recently become a national priority. Surveys conducted by the National Science Foundation showed that on average Americans could only correctly answer five out of ten questions concerning scientific knowledge. When asked to describe in their own words what it means to study something scientifically, only two percent understood that it was a process of developing and testing theories and hypotheses. Nearly two-thirds of adults were unable to describe the scientific process even in a broad sense. This means that when articles are published concerning new technologies such as medications, medical procedures, or genetics only a small fraction of readers “understand scientific inquiry well enough to assess whether the findings presented in the media have any basis in science.” [1]

Our workforce has a great need for technological expertise in order to boost productivity and develop new innovations that drive economic growth, so America needs to produce as many curious, scientific minds as possible. For the most part, the only exposure people get to the pure sciences is in high school or from introductory science courses in college. If students are confused about the material presented in these courses, it is very unlikely that they would choose a career in these fields.

Physics is notorious for being one of the most difficult subjects for students to understand. Physics education research is constantly probing students' minds in order to better understand what it is exactly that makes physics so difficult and to develop new teaching methods that optimize learning. Most research in this area is focused on introductory level courses. If students do not comprehend basic mechanics concepts (Newtonian concepts) that are taught in introductory physics courses, difficulty in more advanced courses is almost a certainty.

Ibrahim Halloun and David Hestenes suggest that the difficulty students have in learning Newtonian concepts arises from the fact that before they set foot in any physics class, they already have fixed common sense beliefs learned from everyday experiences [2]. However, these “common sense beliefs about motion are generally incompatible with Newtonian theory.”[2] “Consequently, there is a tendency for students to systematically misinterpret material in introductory physics courses.”[2] Naturally, instructors want to direct their teaching methods in order to drive these misconceptions away, but this is not as easy as it sounds.

It has been shown that conventional physics instruction does very little to alter preconceived notions. A “physics diagnostic test” focused on conceptual knowledge was

created by Halloun and Hestenes in order to ascertain this result [2]. The initial test given to students started out as free response so that the most common misconceptions could be incorporated into a multiple-choice version. Various versions of this test were administered to over 1000 introductory mechanics students at the University of Arizona. Once the multiple-choice version was devised, precautions were taken to ensure the validity and reliability of the test. Numerous physics professors and graduate students examined the test, and their suggestions were incorporated into the final version. Eleven graduate students took the test, and they all agreed upon the correct answer to each problem. Twenty-two introductory physics students who took the test were interviewed to confirm that they understood the problem and the answer choices. Lastly, the answers of 31 students who received an A in the class were examined for evidence of misunderstandings in the statements of the problems, but none were found. The resulting diagnostic test was the first version of what is now known as the Force Concept Inventory (FCI). The current version now used all over the country has changed very little from the initial diagnostic test. [2]

Since David Hestenes, Malcolm Wells, and Greg Swackhamer introduced the Force Concept Inventory in 1992, it has been the most widely used tool for evaluating student comprehension of basic Newtonian concepts. The concepts tested on the FCI can be broken down into six categories: Kinematics, Newton's First Law, Newton's Second Law, Newton's Third Law, Superposition Principle, and Kinds of Force [3]. This 30 question multiple-choice test challenges students to answer correctly with the one Newtonian choice over four common sense misconceptions. To provide an example, the first problem on the FCI is given below:

Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant. The time it takes the balls to reach the ground below will be

1. about half as long for the heavier ball as for the lighter one.
2. about half as long for the lighter ball as for the heavier one.
3. about the same for both balls
4. considerably less for the heavier ball, but not necessarily half as long.
5. considerably less for the lighter ball, but not necessarily half as long.

Because of its design, the FCI can help instructors identify which misconceptions their students hold. After the misconceptions have been identified, instructors can then develop a plan of action to attack them. It is also common practice to give the FCI on the first day of class (pre-test) and then again at the end of the semester (post-test) to evaluate an instructor's teaching method. This is what Halloun and Hestenes did to confirm that conventional teaching methods created little change in the pre-test and post-test scores. They also found that the pre-test score for an individual provided a good prediction of that student's performance in the class. Because of this, the authors suggest that the FCI can be used as a placement test.

The FCI is a very powerful tool. However, the surprisingly poor performance by students on the test has led instructors and researchers to question the validity of the FCI. Specifically, numerous studies have been conducted to determine the following questions: Should the FCI really be used as a placement test? Is the FCI gender biased? Are the wrong answer choices the most common misconceptions that students hold? What are the effects of changing the FCI to a free response test? What is the FCI really measuring? And is the test context sensitive? Even though some of these studies indicate that the FCI is a less than perfect test, these results are not without dispute, and those who believe the test has faults will still admit that the FCI is a powerful instrument.

A study at the University of Minnesota concluded that the FCI should not be used as a placement exam because even though high scores on the pre-test correspond to those students who do well in their introductory calculus-based physics class, the pre-test does not do a good job of predicting failure in the class. A majority of students who performed poorly on the FCI still received grades of A, B, or C. If the FCI were used as a placement test, many students would be ill advised not to take the class. [4]

Most studies investigating the FCI are not attempts to discredit the test. They are only trying to improve it so as to receive the maximum amount of information about how students are thinking. The study of mental modes is one such study. Mental mode refers to the mindset of a student when encountering a specific problem. For example, students use three models when dealing with Newton's second law: the Newtonian model ($F=ma$), the Aristotelian model ($F=mv$), and a hybrid model that is a combination of the other two [5]. Schecker and Gerdes attempted to determine if students consistently selected the wrong answer associated with a particular model, but since all three models were not represented in the answer choices of each question on the FCI, their analysis was inconclusive. However, they did notice that the model applied depends on the context of the question. For example, a student may answer differently on a question about a cannon ball than they would if the question involved a baseball even if the two questions are testing over the same concept. [6]

Contextual effects are again questioned when studying whether or not the FCI is gender biased. It is known that females normally do not perform as well as males on the test, but whether the FCI is to blame is the question. The University of Wisconsin is currently studying the effects of changing some of the questions to a more female-

oriented context. For example, instead of a boy throwing a rock, a girl throws a teddy bear. Students were administered both the FCI and the “gender” test. The preliminary data shows that both men and women did better on the “gender” test than on the FCI. However, the difference in scores between the two tests was larger for women than the gain obtained by men. From this data it would seem that the FCI is gender biased. At the very least, this study could provide evidence for the context sensitivity of the FCI. [7]

Huffman and Heller produced a study to determine what exactly the FCI measures. A factor analysis study was performed to determine how the questions within each of the categories are related. The correlations found indicate that “the questions on the FCI are only loosely related to each other and do not necessarily measure a single force concept or the six conceptual dimensions of the force concept.” As an explanation, it is suggested that students are using bits and pieces of information to understand forces. In this case, the pieces of knowledge used may depend on how familiar the student is with the context of the problem. “Students may be more familiar with hockey pucks than with rockets, and this experience with the context can affect their understanding of the concept.” If this is the case, performance on the FCI may depend on the context of the questions. [8]

In a response to the factor analysis study, Hestenes claims that the study only supports the fact that students are not understanding Newtonian concepts and that professors should have been used in Huffman and Heller’s study rather than students in order to obtain conclusive evidence [9]. Regardless of whether or not the FCI is testing a single force concept, six dimensions of force concept, or some other dispersion of force concept, a valid question has been raised yet again. Is the FCI context sensitive?

This is the main question we wish to examine within this paper. A test was created which contains ten context modified FCI questions. This test will be referred to as the transformed test. Over five semesters, 654 students were given the transformed test as well as the FCI, but we only examined the ten problems that correspond to the transformed test. Approximately half of the students were given the FCI first followed by the transformed test, and the other half took the transformed test first. The responses of each pair of questions were compared to determine whether a student answered consistently correct, consistently wrong, or inconsistently. It should be noted that consistently wrong means not only that the student answered incorrectly on both FCI and transformed questions but that they answered with the set of corresponding wrong answers. By examining this data the following questions will be answered:

1. Is the FCI context sensitive?
2. What portion of student responses is attributable to random answering or uncertainty?
3. What effect, if any, did the order in which the tests were given have on a student's response?

For the FCI to be considered a valid instrument in physics education, it has to be stable to changes in context. A ball being changed to a mass, for example, should not affect how a student answers the question, but if it does, the FCI would then be measuring how much a student knows about a particular situation rather than what he or she knows about forces. In this case, the conclusion Huffman and Heller reached, that students are using partial knowledge to answer questions and that their partial knowledge only applies to certain situations, would also be supported. Assuming the FCI is stable to context changes, this study will add support for the FCI's validity.

Knowing what portion of the responses is due to random answering or uncertainty provides a method of determining how many students are missing questions because of consistently applying an incorrect model. This information can be used to decide how current teaching methods need to be changed. Students who answer randomly or with uncertainty may only need more clarification of the concept being tested, but those who are continually applying a wrong model will need more hands-on experience to change their previously held beliefs.

Since some students took the transformed test first and others took the FCI first, the order the tests were given in may affect the way a student responds. We would want to know how taking the first test influences a student's answers to the second test if the order does matter. It is possible that misconceptions are being learned from the first test and consequently, being applied to the second test. If this is the case, then the FCI should only be given in situations where the correct answers can be reviewed with the class.

Chapter 2: Data Collection

Given the impact that the FCI has had on physics education, it is necessary that it be tested in as many ways as possible. Testing the context sensitivity can provide information regarding what the FCI is really testing and how much understanding of forces students have really obtained. The construction of the transformed test allows the determination of context changes that produce a significant effect. The FCI questions that were modified were chosen only because they were easy to change without altering the concept that was being tested.

Six types of transformations were used to modify the FCI. The first involves changing the physical system to another system. For example, “truck” was changed to “bowling ball” and “compact car” was changed to “marble” in question 4. This transformation was applied to problems 4, 12, 28, and 30. Some questions make the system more abstract, like changing “truck” to “mass” (problems 1, 25, and 27). Redundant wrong answers are removed from questions 1, 6, and 12. For example, in problem 6, the correct answer is a straight trajectory, and there are three choices for trajectories curving to the right. The transformation applied removes two of the right curving trajectories. Figures were added or removed from some problems. A figure was added to problem 25, and the figure was removed from question 28. Another transformation involves re-ordering multiple-choice answers (problems 1 and 28). The last transformation restructures a group of problems. Problems 25 and 27 were removed from a group of three questions, and problems 22 and 24 were removed from a group of four questions. In the introduction, problem 1 was given to provide an example of an FCI question. This problem was transformed by making the system abstract, removing redundant wrong answers, and re-ordering the answers. It appeared on the transformed test in the following form.

Two spheres are the same size but the mass of one is twice as much as the other. The spheres are dropped from a height of 5m at the same time. The time it takes the spheres to reach the ground will be

1. about the same for both spheres.
2. considerably less for the heavier sphere.
3. considerably less for the lighter sphere.

Most problems had multiple transformations applied to insure that the question was not recognizable to the students. Initially, there were 13 problems on the transformed test,

but problems 19, 21, and 29 (problems 11, 13, and 5 on the transformed test) had been changed to free response. These questions are not considered in this research. The Appendix gives the FCI questions used and the corresponding transformed problems.

A short version of the FCI and the transformed test were given at the University of Arkansas to students enrolled in the University Physics I course. The tests were administered at the end of the semester for five semesters ranging from the spring of 2000 to the fall of 2003. The tests were administered during lab time, and the students were allowed 30 minutes to complete the tests. The first test had to be turned in before the second was handed out. This was to insure that students would not look back to the first test and change answers because of the second test. Certain sections of the labs took the FCI first while others took the transformed test first. The order that the tests were given in was only monitored for three of the five semesters, so only these three semesters are being referred to when order is considered.

Students were told that they could earn up to five bonus points for correct responses, but they would not be penalized for incorrect answers. It is a common concern for studies similar to this that students are not trying their best. The opportunity for bonus provides a little added incentive. Those students who left more than two questions unanswered were eliminated from the study. There were very few people who did this, and no evidence was found that students were answering using patterns (like answering all C's or ABCABC...). Once all the data had been collected, a spreadsheet in Excel was created to easily compare responses for each set of corresponding questions.

Chapter 3: Context Sensitivity

To determine if the FCI is stable under context changes, a statistical t-test was applied. This test determines if the FCI and the transformed test can be considered statistically different. A t-test produces the probability that the two tests are statistically equivalent. There is a significant context effect when there is a very low probability of being considered the same. To be a little more specific, you make the hypothesis that the difference in means (averages) is zero. Then by calculating the t-statistic, you can find the probability that the measured mean was observed given the hypothesis that the means are equal.

In general, the t-test assesses whether the means of two populations are statistically different from each other. This analysis is appropriate for our situation because we want to compare the means of the FCI and the transformed test. The t-test is especially helpful because it automatically takes into account the variability of the test scores. It does this by the calculation of the t-statistic, which is a ratio of the difference in means to the measure of variability or dispersion of the scores called the standard error of the difference. For our situation, we actually calculate the difference in averages and test whether the difference is consistent with zero. The t-statistic is positive if the first mean is larger than the second mean and negative if it is smaller. To establish that the ratio is large enough to say that the difference between the groups is not likely to have been chance, a risk level called an alpha level must be set and the number of degrees of freedom must be calculated. The commonly accepted alpha level is 0.05, which means that only five percent of the time would you find a significant difference even if there actually was none. Once this information is gathered, a “table of significance” is

consulted to determine how large the t-statistic must be to indicate a significant difference in the populations.

Since there were 654 people taking both tests, these calculations would have been quite tedious, so a program called Statistical Analysis System (SAS) was used to perform the t-test. The results are given in table 1 where the ‘Mean T’ (transformed average) and ‘Mean FCI’ (FCI average) given for each problem is the percentage of students who answered the problem correctly, and the total under these columns gives the average number of problems correct (out of ten questions). The probability column gives the probability that the two tests are not statistically different (that is, they function the same), so for low probabilities the context changes had a significant effect.

Table 1: t-test Results

T#	FCI #	Mean T	Mean FCI	Diff Mean(T-FCI)	t-statistic	Probability	Change System	Make System Abstract	Remove Redundant Wrong Answers	Add or Remove Figure	Re-order Answers	Restructure Group
1	28	59	55	4.3	2.90	0.0038	x			x	x	
2	22	74	64	9.7	5.68	<0.0001						x
3	24	85	85	0.2	0.12	0.9043						x
4	4	43	42	1.6	0.96	0.3363	x					
6	6	86	86	0	0.00	1.0000			x			
7	12	92	87	5.4	3.93	<0.0001	x		x			
8	1	92	92	0	0.00	1.0000		x	x		x	
9	25	34	37	-3.2	-1.98	0.0481		x		x		
10	27	78	74	3.6	2.34	0.0194		x				x
12	30	65	60	5.1	3.38	0.0008	x					
Total		7.1	6.8	0.25	5.08	<0.001	75	66	33	100	50	66

Overall, the probability that the two tests are not statistically different is less than 0.1%. It can safely be concluded from this that our context changes had a significant effect. However, the overall difference in means is only a quarter of a problem. This means that even though there is some context effect, it makes very little difference in terms of student performance.

For the individual problems, six out of the ten questions showed a significant context effect. Transformed numbers 3, 6, and 8 showed little or no context effects. Problem 4 is questionable because there is a 34% chance that there was no context effect, which is a little too large to safely say there definitely was an effect.

The last six columns represent the transformations applied to each problem. An 'x' in the problem row indicates that the transformation was applied to that problem. The total row beneath these columns gives the percentage of problems where the particular transformation was applied that showed a context effect. For example, four problems had changes to the physical system, and three of those problems showed significant context effects. So that transformation showed a significant context effect 75% of the time. For these percentages to be improved, more problems would need to have the transformation applied, but they do give some idea as to the effects of each transformation.

Making the system abstract and restructuring a group of problems produced a context effect for two of the three problems. Removing redundant wrong answers was only effective for one of the three problems, and for that one problem, it is likely that the other transformation applied, changing the system, was the reason for the context effect rather than removing redundant wrong answers. This is likely for two reasons. The other two problems where wrong answers were removed had absolutely no difference in the

means, and changing the physical system was an effective transformation for most of the problems where it was applied. Because only two problems were tested for adding/removing figures and re-ordering answers, the percentages given may or may not give a good indication as to the effectiveness of the transformation.

It has been established that context changes do have a significant effect on student responses, but since the effect is not large enough to create dramatic changes in the overall averages, it cannot be said that the FCI is an invalid test. Instead, the result implies that students do not have a deep enough understanding of forces to be able to consistently apply their knowledge in a variety of situations. This problem needs to be corrected. In order to address this issue, it is first necessary to analyze exactly how students are answering.

Chapter 4: Correlation

Even though the context changes did have an effect on test responses, it may be possible to predict how a student will perform on one test based on the responses given on the other test. The correlation coefficient determines how well two variables are linearly related, so it describes how accurate a prediction of one variable based on the other variable could be. The correlation coefficient ranges from negative one to positive one. The closer it is to zero, the less related the variables are. If the correlation is close to one or negative one, then the variables are highly correlated, and the prediction will be a good one.

Using the transformed test as one variable and the FCI as another variable, the correlation coefficient, R , for each problem was calculated. They are given in table 2.

For most individual problems R is of a mid-range value meaning the two problems are related, but predicting the answer to one question based on the answer to the corresponding question would not be very accurate. However, for the entire tests, the correlation is 0.82, which is very close to 1. It can be concluded from this that a good approximation of one test score based on the other test score can be made.

Table 2: Correlation

T#	FCI#	R	R^2
1	28	0.71	0.49
2	22	0.55	0.31
3	24	0.59	0.35
4	4	0.66	0.44
6	6	0.78	0.61
7	12	0.34	0.12
8	1	0.70	0.49
9	25	0.62	0.38
10	27	0.58	0.34
12	30	0.68	0.46
Total		0.82	0.67

The squared correlation coefficient (R^2) is the proportion of variance in the data that is accounted for by the model that the two variables are equal. Mathematically, R^2 is equal to the ratio of predicted variance (s^2_{pred}) to the total variance (s^2_{total}) where $s^2_{\text{total}} = s^2_{\text{pred}} + s^2_{\text{error}}$. The error variance, s^2_{error} , is the variance left in the data after the model, $T = \text{FCI}$, is applied. With some manipulation, the R^2 ratio can be rewritten as $R^2 = 1 - (s^2_{\text{error}} / s^2_{\text{total}})$. The smaller the error variance is, the closer R^2 is to one and consequently, the better the prediction of the dependent variable will be.

For a majority of the individual problems, R^2 is very low. A substantial amount of variance remains after the hypothesis that the two responses should be equal is accounted for, and overall 33% of the variance cannot be accounted for by the

relationship between the tests. This remaining variance in the responses must then be due to uncertainty or randomness in the students' responses.

Chapter 5: Response Distribution

Upon examining the responses for both the FCI and transformed test, each question pair was labeled as consistently correct (CC), consistently wrong (CW), or inconsistent (I). To be consistently correct, both answers on the FCI and the transformed test had to be correct. Consistently wrong means that both answers given were incorrect and that the wrong answers were equivalent. In other words, the student consistently applied an incorrect model. A pair was labeled inconsistent if the answer on one test was right and the other wrong or if both were wrong but for different reasons. Table 3 gives the number of consistently correct pairs (NCC), the number of consistently wrong pairs (NCW), and the number of inconsistent pairs (NI) as well as the matching percentages for each.

Table 3

F#	T#	NCC	NCW	NI	%CC	%CW	%I
28	1	326	186	142	49.8	28.4	21.7
22	2	387	98	169	59.2	15.0	25.8
24	3	521	52	81	79.7	8.0	12.4
4	4	223	290	141	34.1	44.3	21.6
6	6	547	75	32	83.6	11.5	4.9
12	7	549	18	87	83.9	2.8	13.3
1	8	584	30	40	89.3	4.6	6.1
25	9	179	288	187	27.4	44.0	28.6
27	10	452	89	113	69.1	13.6	17.3
30	12	359	162	133	54.9	24.8	20.3
All	All	4127	1288	1125	63.1	19.7	17.2

The percentage of consistently wrong responses is worth discussing here because it provides insight into strongly held misconceptions. Problem 4 has the highest

percentage of consistently wrong answers at 44.3%. The FCI question reads, “A large truck collides head-on with a small compact car. During the collision...” It then gives five answer choices that compare the force the truck exerts on the car and the force the car exerts on the truck, which are actually equal. However, most students said that the truck exerts the greater force. They were applying the misconception that greater mass always means larger force. This is the same misconception that was most commonly applied to problem 1, which also has a high percentage of consistently wrong responses.

Problem 9 was answered consistently wrong 44% of the time. It reads, “A woman exerts a constant horizontal force on a large box. As a result, the box moves across a horizontal floor at a constant speed v_0 . The constant horizontal force applied by the woman...” which should be completed with “has the same magnitude as the total force that resists the motion of the box.” But students most often answered that the woman’s force is greater than the force that resists the motion of the box. Here they are implementing the misconception that motion/velocity implies force. This same misconception is applied to problem 12 even though it is a different situation.

Overall, 17.2% of the population’s responses were inconsistent, so apparently many students were changing their answers from one test to another. For individual questions the inconsistency ranges from as low as 4.9% to as high as 28.6%. It is easy to notice that those questions with the highest level of inconsistency also had the highest levels of consistently wrong answers and the lowest levels of consistently correct responses. For teaching purposes, it would be more than helpful to know why the students are answering as they do. Is the inconsistency due to students answering randomly because they cannot answer the question in any context, or is it due to students

being uncertain of their knowledge? We also want to know what percent of the wrong answers is the result of using an incorrect model as opposed to answering with uncertainty.

Chapter 6: Random Model

A model was created to predict what portion of responses is due to students answering randomly as opposed to students who answer either with correct knowledge or incorrect knowledge. This was done by figuring how many combinations of answers were consistently correct, consistently wrong, and inconsistent as described in the previous chapter. For most problems there were five answer choices. Only one combination of an FCI answer and transformed answer will be consistently correct. Four possible pairs will yield a consistently wrong response, and the other twenty possibilities are inconsistent. Using these numbers, NCC, NCW, and NI can be described by the following equations:

$$NCC = N \cdot P_k + (1/25)N \cdot P_r$$

$$NCW = N \cdot P_w + (4/25)N \cdot P_r$$

$$NI = (20/25)P_r$$

In these equations, P_r is the probability that a student is answering randomly, P_k is the probability that a student is answering with correct knowledge, and P_w is the probability that a student is answering based on incorrect knowledge. Solving for these three variables gives:

$$P_k = (NCC/N) - (1/25)P_r$$

$$P_w = (NCW/N) - (4/25)P_r$$

$$P_r = (25/20)(NI/N).$$

The appropriate corrections were made to the fractions for problems 6, 7, and 8 since they had different numbers for combinations of responses. The calculated probabilities for each question are given in Table 4.

Table 4: Random Model

T#	N	NCC	NCW	NI	P_r	P_k	P_w	%Bad Model
1	654	326	186	142	0.27	0.49	0.24	47
2	654	387	98	169	0.32	0.58	0.10	24
3	654	521	52	81	0.16	0.79	0.06	27
4	654	223	290	141	0.27	0.33	0.40	58
6	654	547	75	32	0.06	0.83	0.11	65
7	654	549	18	87	0.17	0.83	0.00	0
8	654	584	30	40	0.09	0.89	0.02	18
9	654	179	288	187	0.36	0.26	0.38	51
10	654	452	89	113	0.22	0.68	0.10	31
12	654	359	162	133	0.25	0.54	0.21	46
All	6540	4127	1288	1125	0.22	0.62	0.16	42

The bad model column gives the percentage of students not answering consistently correct that can be attributed to the consistent use of an incorrect model. Taking the consistently wrong probability and dividing it by the sum of the random probability and the consistently wrong probability calculated this (%Bad Model = $P_w/(P_w+P_r)$). Overall this model predicts that 42% of the students who answered incorrectly did so because of a strongly held misconception.

This model has one downfall. It cannot accurately predict the dispersion of inconsistent answers. Inconsistent pairs of responses come in two forms: both responses being incorrect and one response correct while the other response is wrong. NIw gives the number of students who answered both questions inconsistently wrong, and NIc

denotes the number of students who answered one question correct but the other wrong. The model inherently predicts these values from the equation $NI=(20/25)N \cdot Pr$. Since there are eight ways to answer and have one problem right while the other is wrong, NIc is given by: $NIc=(8/25)N \cdot Pr$. The other 12 combinations (of the 20 inconsistent combinations) are possibilities for answering inconsistently wrong, so $NIw=(12/25)N \cdot Pr$. These predicted values dramatically overestimate NIw and underestimate NIc .

Table 5: Predictions Based on Model

T#	Predicted NIc	Actual NIc	Predicted NIw	Actual NIw
1	57	96	85	46
2	67	131	100	38
3	33	67	50	14
4	57	108	85	33
6	13	32	19	0
7	36	78	53	9
8	19	31	28	9
9	75	110	113	77
10	46	95	69	18
12	52	102	78	31
all	460	850	691	275

Chapter 7: Uncertain Model

Since the random model does not predict the dispersion of the inconsistent answers very well, a different model was created that separates students into three categories: those answering with correct knowledge, those answering with incorrect knowledge, and those answering with uncertainty. The same method of finding the number of answer combinations that are consistently correct, consistently wrong, or inconsistent is applied except here the inconsistent responses are broken down into two categories. As mentioned previously, there are twenty possibilities for inconsistent responses, but only four of those are possibilities of getting the question right on the

transformed test and wrong on the FCI. Combinations where the student would get the FCI question correct and the transformed question wrong provide another four possibilities. The other twelve possibilities make up the combinations of answers where both responses were wrong but not consistently wrong.

In order to calculate the uncertain probability (P_u), consistently wrong probability (P_w), and consistently correct probability (P_k), a fourth probability must be introduced. It is the probability of getting one problem correct, denoted P_c . So the probability of answering wrong on one problem is $1-P_c$. Then the probability of answering consistently correct is P_c^2 , and the probability of answering one question right and one question wrong is $P_c(1-P_c)$. There are 16 possibilities of getting both questions wrong. Four of these are consistently wrong, and the other 12 are inconsistently wrong. Then the probability of answering consistently wrong is $(4/16)(1-P_c)^2$, and the probability of answering inconsistently wrong is $(12/16)(1-P_c)^2$.

Using these probabilities, NCC, NCW, NIw, and NIc can be represented according the following equations:

$$NCC = N \cdot P_k + N \cdot P_u \cdot P_c^2$$

$$NCW = N \cdot P_w + N \cdot P_u \cdot (1/4)(1-P_c)^2$$

$$NIw = N \cdot P_u \cdot (3/4)(1-P_c)^2$$

$$NIc = 2 \cdot N \cdot P_u \cdot P_c \cdot (1-P_c).$$

The NIc equation contains a factor of two because there are two forms of answering one question correct and one question incorrect. Either the FCI question is right and the transformed wrong or vice-versa. The last two equations involve only P_u and P_c , so those probabilities are solved for first. Then P_k and P_w can be solved for as well. Using the

collected data, these probabilities are calculated. They are given in table 6. Notice that P_u , P_w , and P_k sum to one except for problem 6. In this problem, P_c is one, so calculating the other probabilities would require dividing by zero.

This model shows that strongly held incorrect models are only used in four of the problems. The problems with the highest probability of students answering based on incorrect knowledge are problems 1, 4, 9, and 12. These are the same problems that were discussed in Chapter 5 because they had the largest amounts of consistently wrong responses.

Table 6: Uncertain Model

T#	NCC	NCW	NI	NI_w	NI_c	P_c	P_u	P_w	P_k	%Bad Model
1	326	186	142	46	96	0.44	0.30	0.26	0.44	46
2	387	98	169	38	131	0.56	0.41	0.13	0.46	24
3	521	52	81	14	67	0.64	0.22	0.07	0.70	24
4	223	290	141	33	108	0.55	0.33	0.43	0.24	57
6	547	75	32	0	32	1	undef	undef	undef	-
7	549	18	87	9	78	0.76	0.33	0.02	0.65	6
8	584	30	40	9	31	0.56	0.10	0.04	0.86	29
9	179	288	187	77	110	0.35	0.37	0.40	0.23	52
10	452	89	113	18	95	0.66	0.33	0.13	0.55	28
12	359	162	133	31	102	0.55	0.32	0.23	0.45	42
All	4127	1288	1125	275	850	0.54	0.26	0.18	0.56	41

Here the numbers are very close to those given by the random model, which means that the random model is not a bad model. It just should not be used to predict the dispersion of inconsistent responses. The uncertain model produces the correct dispersion of inconsistent responses. The dispersion of inconsistent responses is important because it classifies two types of students. Those students who answer inconsistently but still get one of the questions correct (given by NI_c) seemingly have some idea about what they are doing. Either these students are not confident enough in

their knowledge to consistently apply it, or they do not completely understand the concept being tested. Since NI_c is significantly large, this provides further evidence of context sensitivity.

The other type of student answers both questions wrong and the two wrong answers do not correspond. These students know very little about the concept being tested. The random model predicts many more of the second type of student than is actually seen in the data, but the uncertain model gives the correct numbers for each (based on the collected data). Being able to predict the number of students who have different degrees of knowledge is helpful for instructors to determine what they are up against.

Chapter 8: Testing Orders

When determining why the students are answering as they do, the order in which the tests were given must be considered. Since some students were given the transformed test first and others were given the FCI first, we need to know if the order the tests were given in had an effect on student responses. If the order does matter, then it can be concluded that the first test given caused certain responses on the second test. Specifically, it may be possible to conclude that misconceptions given on the first test may be transferring to the second test for a particular order. To determine if the order has a significant effect, a statistical F-test is applied.

The F-test determines whether two samples drawn from different populations have the same variance. This test is very similar to the t-test, except here we test the hypothesis that the two standard deviations are equal. The F-statistic is calculated by

dividing the variance of one sample by the variance of the other sample. The closer the F-statistic is to one, the stronger the evidence for equal variances. The probability resulting from the F-statistic tells with what certainty the hypothesis can be accepted. Low probabilities indicate that the standard deviations are statistically different.

Three F-tests were conducted. The first compares the transformed test for one testing order against the transformed test for the other order. The second F-test is applied to the FCI scores rather than the transformed scores. The last F-test determines whether the difference in transformed and FCI scores for one order is significantly different from the difference in scores for the other order. Table 7 gives the F-statistics and probabilities for each of the tests.

Table 7: F-test Results

T#	FCI#	T(F)	T Prob	FCI(F)	FCI Prob	Diff(F)	Diff Prob
1	11	0.73	0.3932	2.74	0.0987	1.52	0.2192
2	7	0.24	0.6258	3.27	0.0715	5.98	0.0149
3	8	0.89	0.3470	0.01	0.9368	1.13	0.2880
4	2	3.11	0.0789	2.85	0.0921	0.01	0.9368
6	3	0.01	0.9372	0.19	0.6653	0.47	0.4915
7	4	0.47	0.4954	0.00	0.9675	0.30	0.5814
8	1	0.24	0.6216	2.26	0.1338	5.88	0.0158
9	9	4.32	0.0383	0.01	0.9362	6.23	0.0130
10	10	0.34	0.5578	1.43	0.2321	0.59	0.4425
12	13	0.13	0.7194	4.18	0.0416	6.57	0.0108
Total		0.04	0.84	3.82	0.05	11.19	0.0009

For the transformed test, there is an 84% chance that the order in which the tests were given made no difference. Consequently, there is a 16% chance that the order does matter. This is very small compared to the 95% chance that the order does matter for the FCI. The result that is most interesting is that there is near a 100% chance that order

matters for the difference of the two tests. It is safe to conclude that overall, the order in which the tests were given had a statistically significant effect on student responses.

We would like to understand how taking the first test changes students' responses on the second test. It is possible that students are learning misconceptions from the first test and applying them to the second test. If this is the case, then there should be a difference in the percentage of consistently wrong answers for the two testing orders. If students are learning misconceptions from the FCI and applying them to the transformed test, then it could be concluded that the FCI should only be given in situations where the correct answers can be reviewed with the class.

Table 8: Order 1-Order 2

T#	%I	%CC	%CW
1	5	7	-12
2	2	0	-1
3	1	-4	3
4	-10	15	-6
6	1	1	-1
7	1	1	-1
8	-1	1	0
9	10	1	2
10	-5	-10	-1
12	0	6	-5
All	1	3	-2

Table 8 gives the difference in percentages of inconsistent responses, consistently correct responses, and consistently wrong responses. Order 1 refers to the order where the transformed test was given first. In order 2, the FCI was given first. The negative percentages indicate that the percentage was higher when the FCI was given first. With the exception of three problems, the consistently wrong percentage was greater when the FCI was given first, but overall this difference is only 2%, which is hardly enough to claim that students are learning misconceptions from the FCI. The overall difference in

inconsistency and consistently correct percentages is also very small, but both are slightly higher when the transformed test is given first.

Chapter 9: Summary and Conclusions

The FCI has been a subject of debate for physics education researchers since it was introduced. Over the years, the FCI has stood up surprisingly well to studies that attempt to pinpoint the cause of low scores on the test. In this study, the context sensitivity of the FCI is analyzed by comparing a context modified test to the FCI, and it is shown that the FCI is slightly context sensitive. However, the context effects are not large enough to cause dramatic changes in the scores of the tests, implying that the FCI is stable enough to be used as a diagnostic instrument. Students are unable to consistently apply their knowledge of forces in a variety of situations. Since the FCI cannot be proved to be the cause of low scores, the cause must then be the students' knowledge. Something must be done to increase understanding of force concepts. Before instructors can evaluate and alter their teaching methods, understanding why students are answering as they do on the FCI is necessary.

A model was created to calculate the probability that a student is answering randomly, the probability that a student is answering with correct knowledge, and the probability that a student is answering based on incorrect knowledge. From these probabilities, the percentage of students who do not answer consistently correct that have a strongly held misconception is found to be 42%. This means that almost half of the student population have strongly held misconceptions. These misconceptions require more hands-on experience so that students can actually see how their thinking was wrong

and what the right way to think about a problem actually is. New lab exercises may be needed to address the concepts that are commonly being misconstrued. For the tested population, there are two misconceptions that are continually applied: motion implies force and greater mass always means larger force.

The previously described model does not correctly predict the dispersion of inconsistent responses, which is why a different model was created. This new model calculates the probability of a student answering with uncertainty, the probability of a student answering based on correct knowledge, and the probability of a student answering using incorrect knowledge. The probabilities found indicate that there is a 26% chance that a student is answering with uncertainty. These students have some idea about the concept being tested, but cannot generalize from one context to another. Correcting this problem is not as difficult as stamping out misconceptions and may only require a little extra instruction to clarify how the concept should be applied in a variety of situations.

In examining why students are responding as they do, the order in which the tests were given must be considered. It is found that the test given first does have some effect on how the second test is answered, but the dispersion of consistently correct answers, consistently wrong answers, and inconsistent responses changes very little between the two orders. So the previous results given by the probability models does not depend on the order the tests were given.

It has been shown over and over again that traditional teaching methods employing “passive-student lectures, recipe labs, and algorithmic-problem exams” produce little or no change in common sense misconceptions [10]. The recent effort to

improve physics education has led to the development of “interactive engagement” (IE) teaching methods, which were designed “to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors.” [10] IE methods produce twice as much improvement in FCI scores than traditional methods do. Considering that this study has shown that a majority of students are still clinging to misconceptions, the use of IE methods would help address this problem. At the very least, students would be more interested in learning physics.

Bibliography


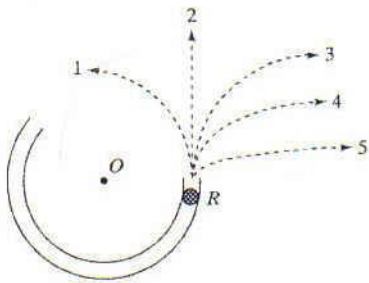

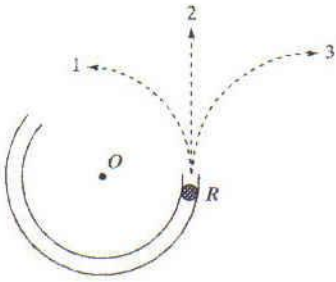
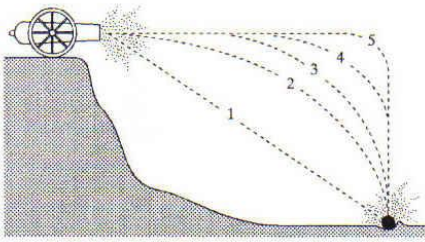
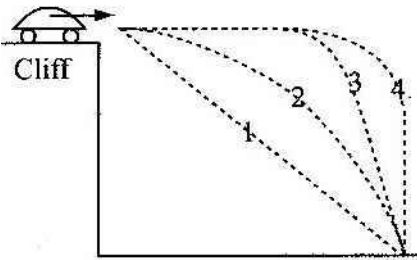
- [1] “Undergraduate Physics Education: Boundary Conditions and Boundless Opportunities.” National Academy of Sciences. Dr. Neal F. Lane.
- [2] Halloun, Ibrahim and David Hestenes. “The Initial Knowledge State of College Physics Students.” *American Journal of Physics*, 53 (1985), 1043-1055.
- [3] Hestenes, David, Gregg Swackhamer, and Malcolm Wells. “Force Concept Inventory.” *The Physics Teacher*, 30 (March 1992), 141-158.
- [4] Henderson, Charles. “Common Concerns About the Force Concept Inventory.” *The Physics Teacher*, 40 (December 2002), 542-547.
- [5] Itza-Ortiz, Salomon, Sanjay Rebello, and Dean Zollman. “Students’ Models of Newton’s Second Law in Mechanics and Electromagnetism.” *European Journal of Physics*, 25 (2004), 81-89.
- [6] Rebello, Sanjay and Dean Zollman. “The Effect of Distractors on Student Performance on the Force Concept Inventory.” *American Journal of Physics*, 72 (2004), 116-125.
- [7] McCullough, Laura. “A Gender Context for the Force Concept Inventory.”
<http://physics.uwstout.edu/staff/mccullough/AAPTJan01San%20Diego.pdf>
- [8] Huffman, Douglas and Patricia Heller. “What Does the Force Concept Inventory Actually Measure.” *The Physics Teacher*, 33 (March 1995), 138-143.
- [9] Halloun, Ibrahim and David Hestenes. “Interpreting the Force Concept Inventory.” *The Physics Teacher*, 33 (1995), 502-506.


- [10] Hake, Richard. "Interactive-Engagement vs Traditional Methods: A Six-thousand-student Survey of Mechanics Test Data for Introductory Physics Courses." *American Journal of Physics*, 66 (1998), 64-74.
- [11] Devore, Jay. Probability and Statistics for Engineering and the Sciences. 6th Ed. Brooks/Cole. 2004.
- [12] Freund, John and Ronald Walpole. Mathematical Statistics. 3rd Ed. Prentice-Hall Inc. 1980.
- [13] Stockburger, David. "Correlation."
<http://www.psychstat.smsu.edu/inrobook/sbk17m.htm>

Appendix

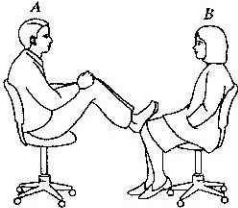
FCI problems and the corresponding transformed problems

FCI	Transformed
<p>Problem 1. Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant. The time it takes the balls to reach the ground below will be</p> <ol style="list-style-type: none"> 5. about half as long for the heavier ball as for the lighter one. 6. about half as long for the lighter ball as for the heavier one. 7. about the same for both balls 8. considerably less for the heavier ball, but not necessarily half as long. 9. considerably less for the lighter ball, but not necessarily half as long. 	<p>Problem 8. Two spheres are the same size but the mass of one is twice as much as the other. The spheres are dropped from a height of 5m at the same time. The time it takes the spheres to reach the ground will be</p> <ol style="list-style-type: none"> 4. about the same for both spheres. 5. considerably less for the heavier sphere. 6. considerably less for the lighter sphere.
<p>Problem 4. A large truck collides head-on with a small compact car. During the collision</p> <ol style="list-style-type: none"> 1. the truck exerts a greater amount of force on the car than the car exerts on the truck. 2. the car exerts a greater amount of force on the truck than the truck exerts on the car. 3. neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck. 4. the truck exerts a force on the car but the car does not exert a force on the truck. 5. the truck exerts the same amount of force on the car as the car exerts on the truck. 	<p>Problem 4. A bowling ball collides with a marble. The bowling ball is much heavier than the marble. During the collision, let the force of the bowling ball on the marble be $F_{\text{b.ball on marble}}$ and the force of the marble on the bowling ball be $F_{\text{marble on b.ball}}$.</p> <p>During the collision</p> <ol style="list-style-type: none"> 1. $F_{\text{b.ball on marble}} = F_{\text{marble on b.ball}}$ 2. $F_{\text{b.ball on marble}} > F_{\text{marble on b.ball}}$ 3. $F_{\text{b.ball on marble}} < F_{\text{marble on b.ball}}$ 4. $F_{\text{b.ball on marble}} = 0, F_{\text{marble on b.ball}} = 0$ 5. $F_{\text{b.ball on marble}} > 0, F_{\text{marble on b.ball}} = 0$

FCI	Transformed
<p data-bbox="345 233 829 302"><i>(The following description and figure is actually given before problem 5)</i></p> <p data-bbox="345 306 846 627">The accompanying figure shows a frictionless channel in the shape of a segment of a circle with its center at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. Forces exerted by the air are negligible. A ball is shot at high speed into the channel at P and exits at R.</p>  <p data-bbox="345 911 805 1089">Problem 6. Which of the paths 1-5 would the ball most closely follow after it exits the channel at R and moves across the frictionless table top?</p> 	<p data-bbox="868 233 1369 554">Problem 6. The figure below shows a frictionless channel in the shape of a segment of a circle with its center at O. The channel has been anchored to a frictionless horizontal table top. You are looking down at the table. Forces exerted by the air are negligible. A ball is shot at high speed into the channel at P and exits at R.</p>  <p data-bbox="868 804 1338 947">Which of the paths 1-3 below would the ball most closely follow after it exits the channel at R and moves across the frictionless table top?</p> 
<p data-bbox="345 1440 834 1619">Problem 12. A ball is fired by a cannon from the top of a cliff as shown below. Which of the paths 1-5 would the cannon ball most closely follow?</p> 	<p data-bbox="868 1440 1369 1583">Problem 7. A car is driven off the top of a cliff at high speed as shown below. Which of the paths 1-4 would the car most closely follow?</p> 

FCI	Transformed
<p><i>(The following description and figure are actually given before problem 21.)</i></p> <p>A spaceship drifts sideways in outer space from point P to point Q as shown below. The spaceship is subject to no outside forces. Starting at position Q, the spaceship's engine is turned on and produces a constant thrust (force on the spaceship) at right angles to the line PQ. The constant thrust is maintained until the spaceship reaches point R in space.</p>  <p>Problem 22. As the spaceship moves from point Q to point R its speed is</p> <ol style="list-style-type: none"> 1. constant. 2. continuously increasing. 3. continuously decreasing. 4. increasing for a while and constant thereafter. 5. constant for a while and decreasing thereafter. 	<p>Problem 2. A spaceship is drifting sideways in outer space. It is subject to no outside forces. It turns on its engine which generates a constant thrust (constant force on the spaceship). While the engine is turned on the spaceship's speed is</p> <ol style="list-style-type: none"> 1. constant. 2. continuously increasing. 3. continuously decreasing. 4. increasing for a while and constant thereafter. 5. constant for a while and decreasing thereafter.
<p><i>(Problem 23 establishes that 'At point R, the spaceship's engine is turned off and the thrust immediately drops to zero.)</i></p> <p>Problem 24. Beyond position R the speed of the spaceship is</p> <ol style="list-style-type: none"> 1. constant. 2. continuously increasing. 3. continuously decreasing. 4. increasing for a while and constant thereafter. 5. constant for a while and decreasing thereafter. 	<p>Problem 3. After some time, the spaceship's engine is question 2 is turned off and the thrust immediately drops to zero. After the engine is turned off, the speed of the spaceship is</p> <ol style="list-style-type: none"> 1. constant. 2. continuously increasing. 3. continuously decreasing. 4. increasing for a while and constant thereafter. 5. constant for a while and decreasing thereafter.

FCI	Transformed
<p>Problem 25. A woman exerts a constant horizontal force on a large box. As a result, the box moves across a horizontal floor at a constant speed v_0. The constant horizontal force applied by the woman</p> <ol style="list-style-type: none"> 1. has the same magnitude as the weight of the box. 2. is greater than the weight of the box. 3. has the same magnitude as the total force that resists the motion of the box. 4. is greater than the total force resists the motion of the box. 5. is greater than either the weight of the box or the total force that resists its motion. 	<p>Problem 9. Block 1 is used to push block 2. Block 1 applies a constant horizontal force to block 2. As a result, the blocks move across a horizontal surface at a constant speed. The surface has friction.</p>  <p>The constant horizontal force applied by block 1 on block 2</p> <ol style="list-style-type: none"> 1. has the same magnitude as the weight of the block 2. 2. is greater than the weight of the block 2. 3. has the same magnitude as the total force that resists the motion of block 2. 4. is greater than the total force that resists the motion of the block 2. 5. is greater than either the weight of block 2 or the total force that resists its motion.
<p>Problem 27. If the woman in question 25 suddenly stops applying a horizontal force to the block, then the block</p> <ol style="list-style-type: none"> 1. immediately comes to a stop. 2. continues moving at a constant speed for a while and then slows to a stop. 3. immediately starts slowing to a stop. 4. continues at a constant speed. 5. increases its speed for a while and then starts slowing to a stop. 	<p>Problem 10. If block 1 in question 9 suddenly disappears along with its force, then block 2</p> <ol style="list-style-type: none"> 1. instantly comes to a stop. 2. continues moving at a constant speed for a while and then slows to a stop. 3. immediately starts slowing to a stop. 4. continues at a constant speed. 5. increases its speed for a while and then starts slowing to a stop.

FCI	Transformed
<p>Problem 28. In the following figure, student A has a mass of 75 kg and student B has a mass of 57 kg. They sit in identical office chairs facing each other. Student A places his bare feet on the knees of student B, as shown. Student A then suddenly pushes outward with his feet, causing both chairs to move.</p>  <p>During the push and while the students are still touching one another.</p> <ol style="list-style-type: none"> 1. neither student exerts a force on the other. 2. student A exerts a force on student B, but B does not exert any force on A. 3. each student exerts a force on the other, but B exerts the larger force. 4. each student exerts a force on the other, but A exerts the larger force. 5. each student exerts the same amount of force on the other. 	<p>Problem 1. Two people stand on a frictionless icy surface. Person 1 has a larger mass than person 2. Person 1 pushes person 2 with his hands causing both people to slide. During the push while persons 1 and 2 are still touching one another,</p> <ol style="list-style-type: none"> 1. neither person exerts a force on the other. 2. each person exerts the same amount of force on the other. 3. person 1 exerts a force on person 2, but 2 does not exert any force on 1. 4. each person exerts a force on the other, but 2 exerts the larger force. 5. each person exerts a force on the other, but 1 exerts the larger force.

FCI	Transformed
<p>Problem 30. Despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court. Consider the following forces:</p> <ul style="list-style-type: none">A. a downward force of gravity.B. a force by the "hit."C. a force exerted by the air. <p>Which of the above forces is (are) acting on the tennis ball after it has left contact with the racquet and before it touches the ground?</p> <ul style="list-style-type: none">1. A only2. A and B3. A and C4. B and C5. A, B, and C	<p>Problem 12. On a very windy day, a baseball player throws a baseball from the outfield to second base. Consider the following forces:</p> <ul style="list-style-type: none">A. a downward force of gravity.B. the force of the "throw."C. a force exerted by the wind. <p>Which of the above forces is (are) acting on the baseball while it is in flight?</p> <ul style="list-style-type: none">1. A only2. A and B3. A and C4. B and C5. A, B, and C